WOMBAT and WOMBAT-PK:

Bioactivity Databases for Lead and Drug Discovery

Marius Olah[1], Ramona Rad[2], Liliana Ostopovici[2], Alina Bora[2], Nicoleta Hadaruga[2], Dan Hadaruga[2], Ramona Moldovan[2], Adriana Fulias[3], Maria Mracec[2] and Tudor I. Oprea[1,4]

[1] Sunset Molecular Discovery LLC
1704 B Llano Street, S-te 140
Santa Fe NM 87505-5140, USA
Email: tudor@sunsetmolecular.com

[2] Romanian Academy, Institute of Chemistry "Coriolan Dragulescu"
Timisoara
Bd. Mihai Viteazul nr. 24
RO-300223 Timisoara, Romania

[3] Faculty of Pharmacy, University of Medicine and Pharmacy "Victor Babes"
Timisoara,
P-ta Eftimie Murgu Nr. 2
RO-300790 Timisoara, Romania

[4] Division of Biocomputing, MSC11 6145
University of New Mexico School of Medicine
1 University of New Mexico
Albuquerque NM 87131-0001, USA
Email: toprea@salud.unm.edu

contribution to

Chemical Biology: From Small Molecules to Systems Biology and Drug Design

Edited by Stuart L. Schreiber, Tarun M. Kapoor and Günther Wess

WILEY-VCH Verlag GmbH & Co. KgaA, Weinheim

ISBN-13: 978-3-527-31150-7

planned release in January 2007

version 1.2, September 2, 2006

Outlook

This chapter highlights the importance of gathering appropriate and accurate information with respect to chemical structures and associated bioactivities, focused on drug discovery. The contents of WOMBAT and WOMBAT-PK are summarized, and examples are given for some of the problems that are encountered when indexing correct biological properties and chemical structures. Two examples for data mining in WOMBAT are given.

Introduction. The WOMBAT databases

The current paradigm for drug discovery allows a relatively short time period, 6-12 months, for the process that modifies an initial active compound – either from high throughput screening (HTS), or from publications and patents – into a well-characterized lead molecule. During this time, project team members have relatively little time to familiarize themselves with 'prior art', i.e., to gather information pertinent to the new biological target, the disease models, as well as active chemotypes on the intended, or related targets. The task of gathering background information related to chemotypes is made easier if one has access to chemical databases such as Chemical Abstracts via SciFinder [1], Beilstein [2] and Spresi [3], or to medicinal chemistry related patent databases such as the MDL Drug Data Report, MDDR [4], the World Drug Index, WDI [5], and Current Patents Fast Alert [6]. Collections of biologically active compounds include Comprehensive Medicinal Chemistry, CMC [7] and DiscoveryGate [8], whereas the PubChem [9] database, part of the Molecular Libraries Initiative (MLI) [10], is more focused on tools for chemical biology. Clinical pharmacokinetics data for marketed drugs is captured in the Physician Desk Reference, PDR [11], whereas DrugBank [12] also captures compounds in clinical trials.

Primary HTS data are captured in PubChem [9], which has author-defined labels for "active" and "inactive" chemical probes. However, most of the other databases listed above do not capture biological endpoints in a simple searchable manner: There are no fields one can query in a quantitative manner to identify what is the target-related activity of a particular compound, or what other measured properties it has. Such information is important if one considers that (a) not all chemotypes indexed in patent databases are indeed active – some are merely patent claims with no factual basis; that (b) not all chemotypes disclosed as active are equally active, or selective for that matter, on the

target(s) of choice; and that (c) not all compounds sharing the same therapeutic indication behave in the same manner with respect to, e.g., side effects.

Table 1. Examples of annotated databases, modified from [16].

Database	Description	Homepage
AurSCOPE	databases containing biological & chemical information relating to a class of drug targets or a pharmaceutical topic of interest	http://www.aureus-pharma.com/
BIDD	bioinformatics databases about drugs, natural products, protein targets, ADME (Absorption, Distribution, Metabolism and Excretion) / Tox and drug-protein binding	http://bidd.nus.edu.sg/
Bioprint	ligand profiling data including target-specific activity, pharmacology and ADME-related properties	http://www.cerep.fr/
Blueprint	resource for biomolecular data focused on public databases for small molecule/domain interactions	http://www.blueprint.org/
ChemBank	database about small molecules and resources for studying their effects on biology	http://chembank.broad.harvard.edu/
Drugmatrix	pharmacological, pathology and gene expression profiles for benchmark drugs	http://www.iconixpharm.com/
GVK Biosciences databases	chemical structures, biological activities, toxicity and pharmacological data for a large number of compounds curated from patents and journals	http://www.gvkbio.com/
*Ki*Bank	database of chemical structures with associated binding affinity (K_i) for given targets	http://kibank.iis.u-tokyo.ac.jp/
Kinase Knowledgebase	captures published information for therapeutically relevant kinases	http://www.eidogen-sertanty.com/
Jubilant Biosys databases	chemical structures, bioactivities, therapeutically relevant databases for a large number of compounds curated from journals & patents	http://www.jubilantbiosys.com/products.htm
Ligand Info	Small-Molecule Meta-Database which compiles various publicly available small molecule databases	http://ligand.info/
MDL Drug Data Report	contains biologically relevant compounds, including launched and candidate drugs, and well-defined derivatives	http://www.mdli.com/
PDSP K_i	public domain resource that provides information related to drugs and their binding properties	http://pdsp.cwru.edu/
PubChem	provides a high volume of information on the biological activities of small molecules; it links chemical structures to other Entrez databases	http://pubchem.ncbi.nlm.nih.gov/
ZINC	online resource of commercially available compounds dedicated to virtual screening practitioners	http://blaster.docking.org/zinc/

Some of these were considered at AstraZeneca R&D Mölndal, Sweden, in May 2001, to initiate a data gathering project centered primarily on the *Journal of Medicinal Chemistry* (JMC), in collaboration with scientists at the Romanian Academy Institute of Chemistry in Timisoara, Romania. The major goal of this project was to capture chemical structures and associated biological activities disclosed in JMC, with an initial goal of 20,000 entries set for the first year. The first version of this database was available at AstraZeneca R&D Mölndal in May 2002; this version contained 21,700 structures (with

duplicates), and 36,738 experimental activities on 324 targets, captured from 837 JMC papers (1996-1999).

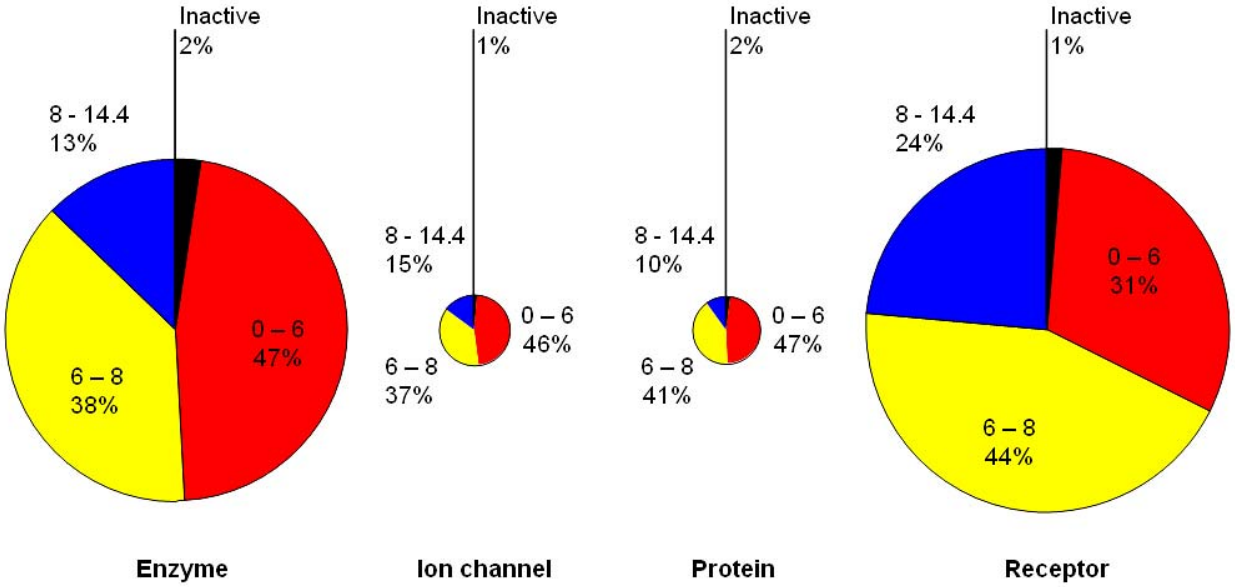


Figure 1. Bioactivity distribution pie charts in WOMBAT 2006.1, classified by target type. The size of the pie chart is proportional to the representation of each target class: enzymes, 42%; ion channels, 7%; proteins 7%; and receptors, 45%.

Because the internal dissemination of this database within AstraZeneca R&D (a company with eleven R&D sites across four continents) was not deemed a success, AstraZeneca decided to discontinue the project as of May 2002. Backed by private funding, the database, renamed WOMBAT (WOrld of Molecular BioAcTivity) in 2003, continued to evolve [13], as discussed for WOMBAT 2006.1, below. Recognizing the paucity of chemical databases that capture clinical pharmacokinetics data in a searchable manner, we further developed WOMBAT-PK (WOMBAT-Pharmacokinetics), to index such data from literature [14]. This chapter summarizes the contents of WOMBAT and WOMBAT-PK [15], some of the problems encountered in appropriately indexing biological activities and correct chemical structures (with focus on machine-readable contents for data mining), and provides some examples of data mining with WOMBAT. Other bioactivity databases [16], focused mostly on patent literature, are shown in Table 1, together with the online references.

WOMBAT 2006.1 Overview

WOMBAT 2006.1 contains 154,236 entries (136,091 unique SMILES, Simplified Molecular Input Line Entry System [17,18]), covering 6,801 series from over 6,791 papers with over 307,700 activities for 1,320 unique targets. All biological activities are automatically converted to the negative $\log_{10}$ of the molar concentration, regardless of activity type. Numerical values for activity are stored in 3 fields; the additional 2 fields capture the experimental error, when reported*. Besides exact numeric values (the vast majority), WOMBAT now captures 'inactives' (3,639), 'less than' (21,926), 'greater than' (635), as well as % inhibition values (8,448 single dose experiments).

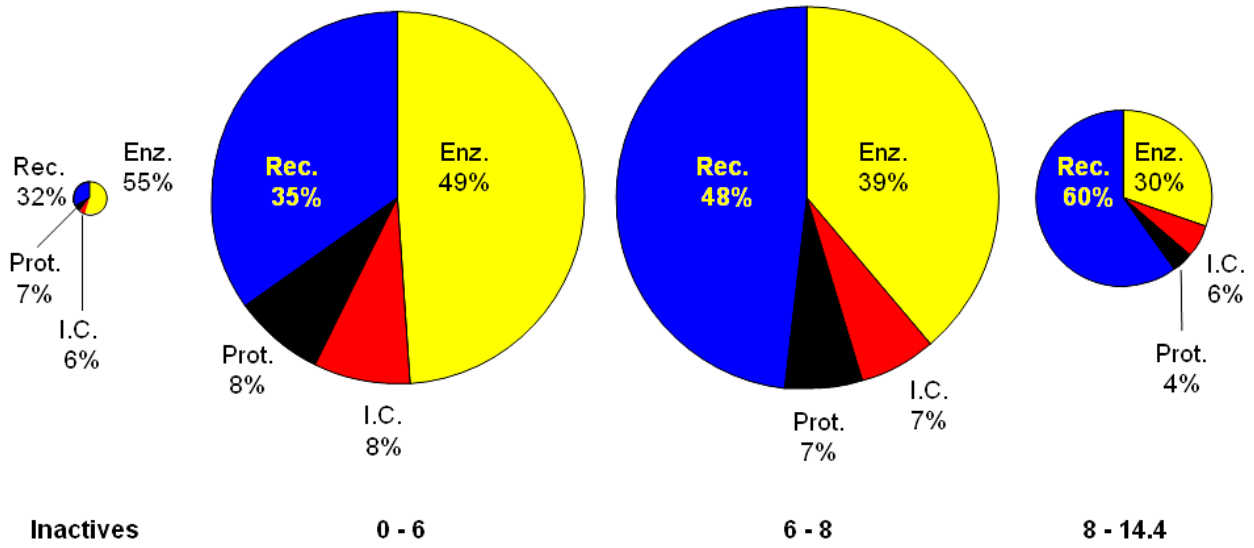


Figure 2. Target type distribution pie charts in WOMBAT 2006.1, classified by activity value (in the $-\log_{10}$ scale). The size of the pie chart is proportional to the representation of each activity category: inactives, 2%, low activity (0 – 6), 18%; medium activity (6 – 8), 41%; and high activity (8 – 14.4), 40%.

The bioactivity distribution by target type is given in Figure 1. Four target types are captured in WOMBAT: *receptors* (which includes GPCRs - G-protein coupled receptors, nuclear hormone receptors, integrins and other receptors, e.g., sigma), *enzymes* (associated with the Enzyme Commission E.C. number [19]), *ion channels* and *proteins* (biological targets that are not known as receptors, enzymes or ion channels, e.g., transporters). A vast majority of the biological activities are related to inhibitors and antagonists: ~56% of the activities are IC_{50} values (and variations), and 37% are K_i values (and variations). Much less frequent are D_2 or EC_{50} values (~3% of the

* In the absence of reported errors, the 3 activity value fields are equal. The decision to index these values for each molecule was taken because 'missing values' are given a different interpretation by statistical techniques.

measurements are for agonists or substrates) and binding affinity constants (~1% K_b and K_d). In WOMBAT 2006.1, enzyme inhibitors populate more of the inactive/low activity bins; while receptor antagonists populate more of the medium/high activity bins (see also Figure 2). The target profile of biological activities is given in Table 2, with focus on some targets classes of current interest to the pharmaceutical industry. Table 2 further indicates the ratio of "actives" in this release of WOMBAT: From this table, one can notice that for some target classes (e.g., phosphatases), there is a relatively small number of "actives", a trend that is observed in most of the indexed enzymes. On the other hand, receptor classes have a higher ratio of "actives". The target type distribution by activity in Figure 2 reflects approximately fifteen years of medicinal chemistry (see also Table 2). Medicinal chemistry publications currently indexed in WOMBAT are listed in Table 3.

Table 2. Target class profile for WOMBAT.2006.1. †

Target Class	Entries	Percent	Actives	Percent
G-protein coupled receptors	50,778	32.92	31,111	20.17
Integrins	3,127	2.03	1,692	1.10
Nuclear Hormone Receptors	4,335	2.81	2,436	1.58
Sigma Receptors	2,123	1.38	1,661	1.08
Ion Channels	13,500	8.75	5,352	3.47
Serine Proteases	7,596	4.92	3,166	2.05
(Oxido)Reductases	7,770	5.04	2,865	1.86
Kinases	9,705	6.29	3,241	2.10
Phosphatases	1,361	0.88	81	0.05
Oxygenases	6,051	3.92	1,716	1.11
Aspartyl Proteases	4,904	3.18	2,881	1.87
Metallo-proteases	4,296	2.79	1,471	0.95
Cysteine Proteases	2,063	1.34	771	0.50
Transporters	5,462	3.54	2,860	1.85
Others	31,165	20.21	N/A	N/A

Table 3. Medicinal chemistry publications covered in WOMBAT.2006.1

Journal Title	Percent	Publication Years
J. Med. Chem.	77.6	1991-2004 [complete] 2005 [partial coverage]
Bioorg. Med. Chem. Lett.	15.4	2002-2003 [complete] 2004 [partial coverage]
Bioorg. Med. Chem.	5.6	2002-2003 [complete] 2004 [partial coverage]
Eur. J. Med. Chem.	1.0	2002-2003 [complete] 2004 [partial coverage]

† Entries indicate the number of structures recorded for each target class, whereas "actives" indicate those entires with an activity of 100 nM or better; percentage values relate to the total number of entries.

The WOMBAT database schemata, illustrated in Figure 3, are further discussed in the next section. Its organization, illustrated in Figures 4-6, shows the 3 panels of the database: The *Bioactivity Summary* panel (Figure 4) provides bioactivity types and values, some basic target information, the minimal reference information as well as structural chemical (2D depiction and SMILES code) and related information (chirality, salt). The *Target and Biological Information* panel (Figure 5) provides detailed target information, including biological information (species, tissue, etc.), detailed target and target class information (including hierarchical classification for G-protein coupled receptors, nuclear hormone receptors and enzymes) as well as further information regarding the bioassays (e.g., radioligand, assay type, etc.). SwissProt [20] reference IDs are stored for most targets (~88%).

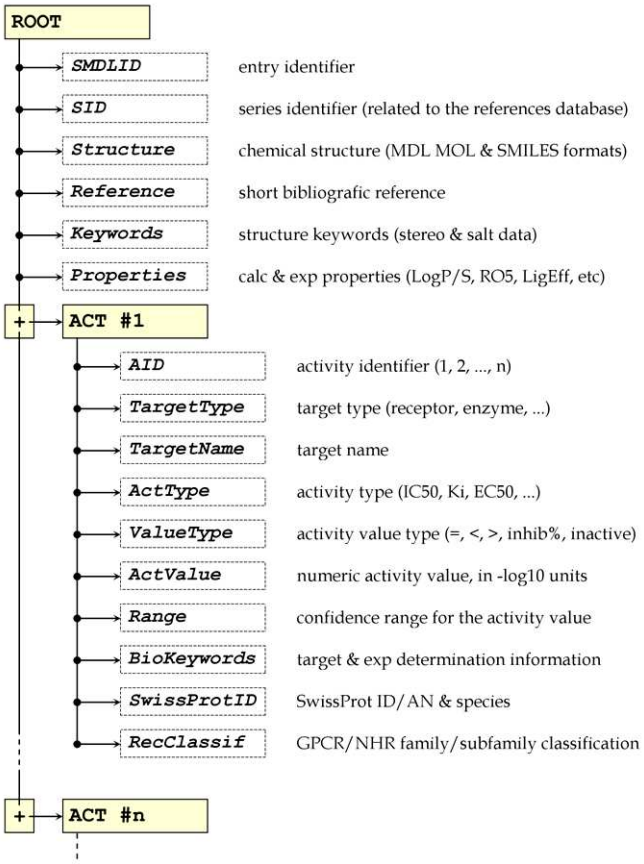


Figure 3. WOMBAT database schemata (simplified).

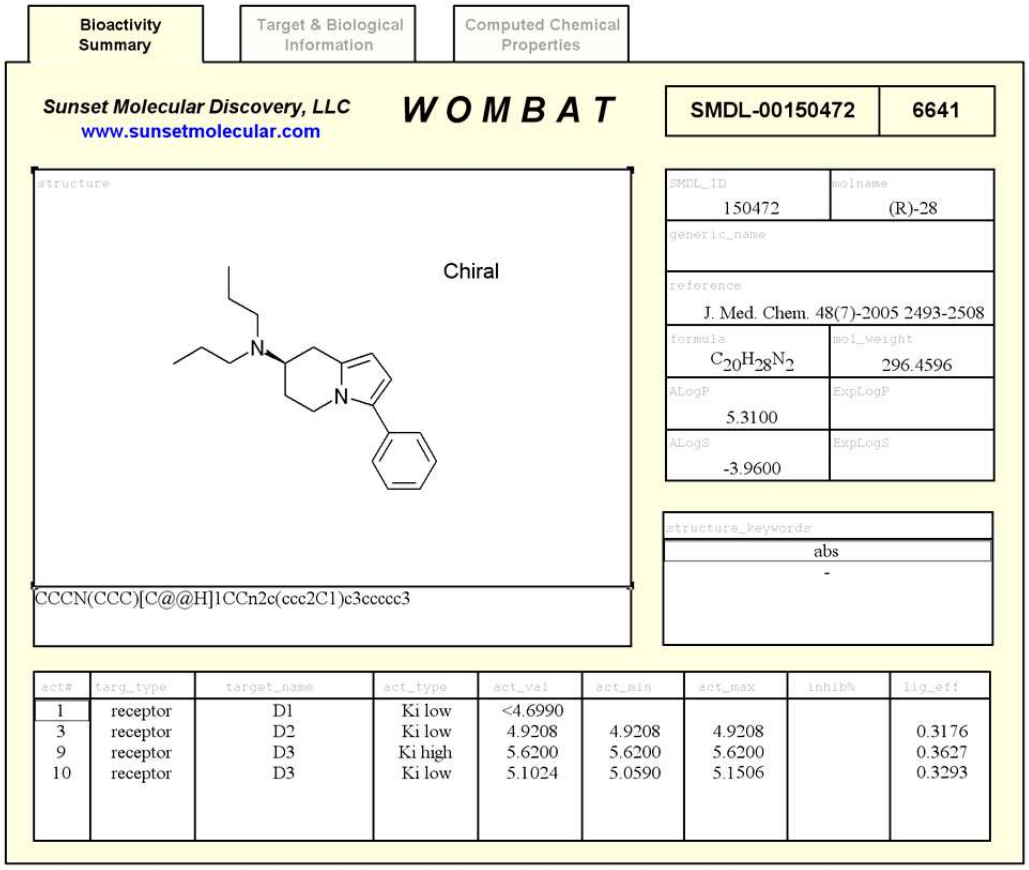


Figure 4. WOMBAT bioactivity summary panel (example).

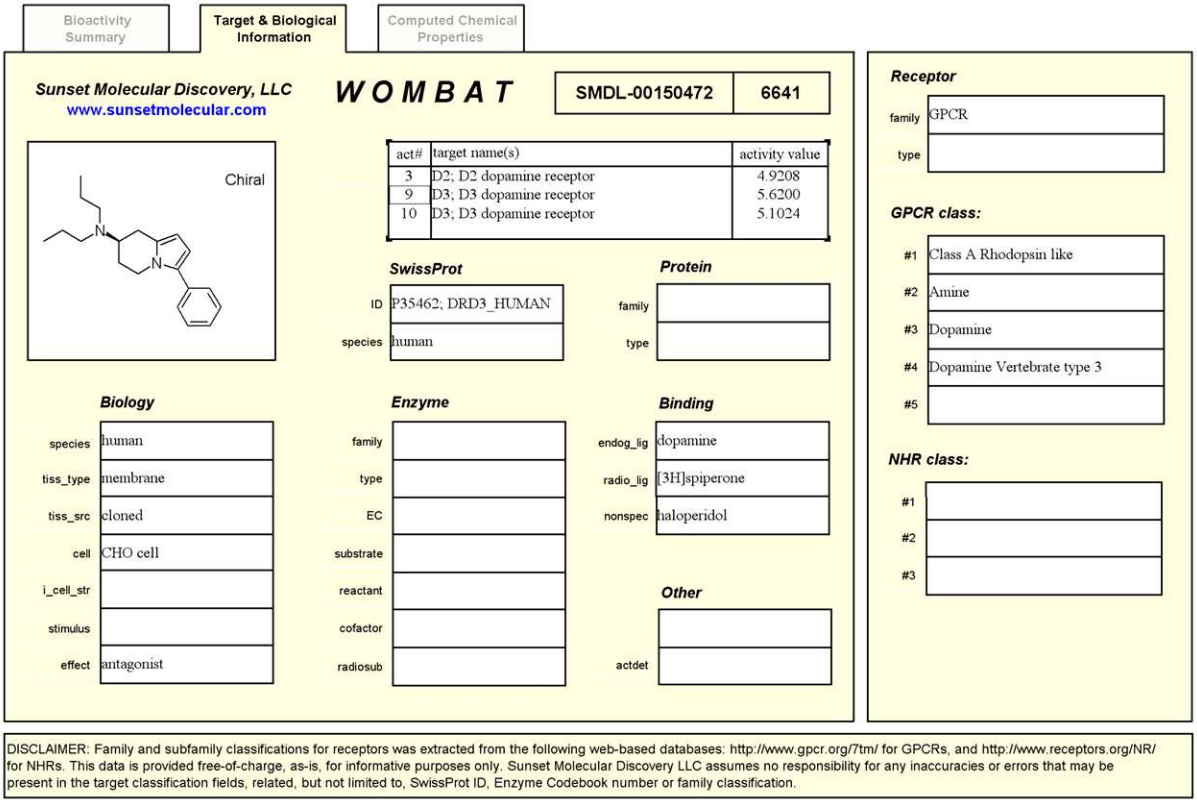


Figure 5. WOMBAT target and biological information panel (example).

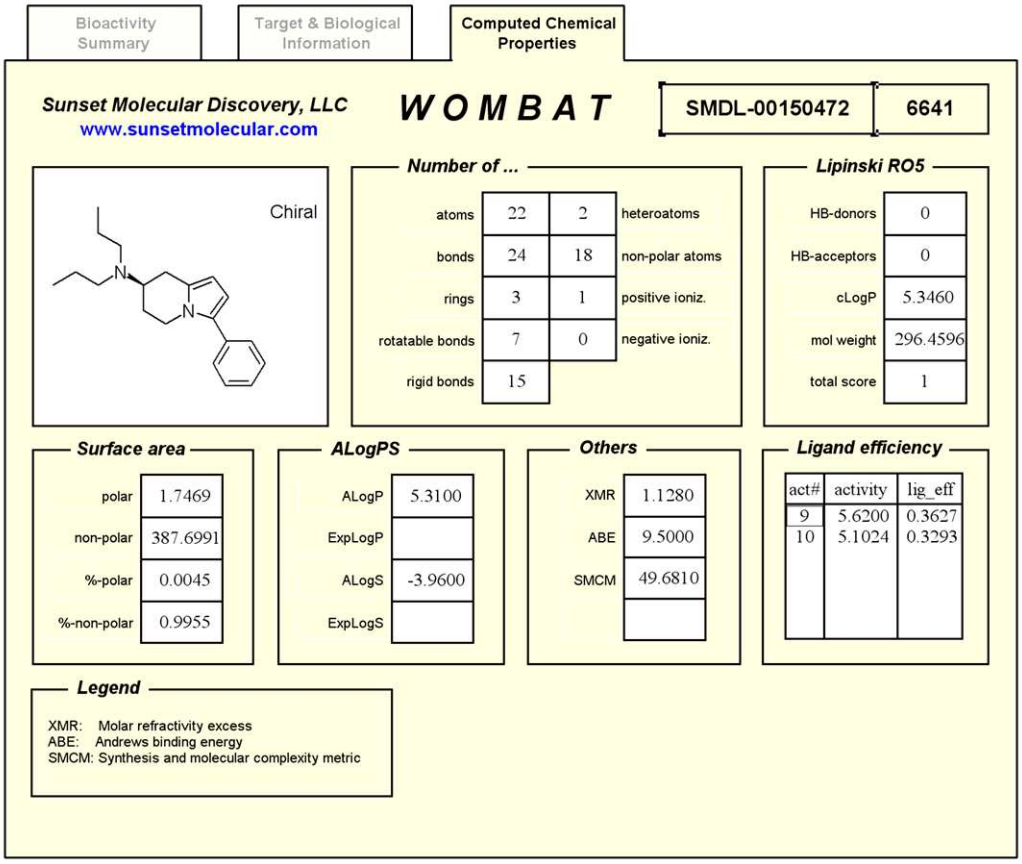


Figure 6. WOMBAT computed chemical properties panel (example).

The *Computed Chemical Properties* panel (Figure 6) includes several calculated and experimental properties for each chemical structure, e.g., counts of miscellaneous atom types, Lipinski's rule-of-five (Ro5) parameters [21], including the calculated octanol/water partition coefficient, ClogP [22] and Tetko's calculated water solubility [23], polar and nonpolar surface areas (PSA, NPSA), etc. Finally, the *Reference Database* contains bibliographic information (Figure 7), including the Digital Object Identifier (DOI) format [24] with URL links to PDF files for all literature entries, as well as the PubMed ID for each paper.

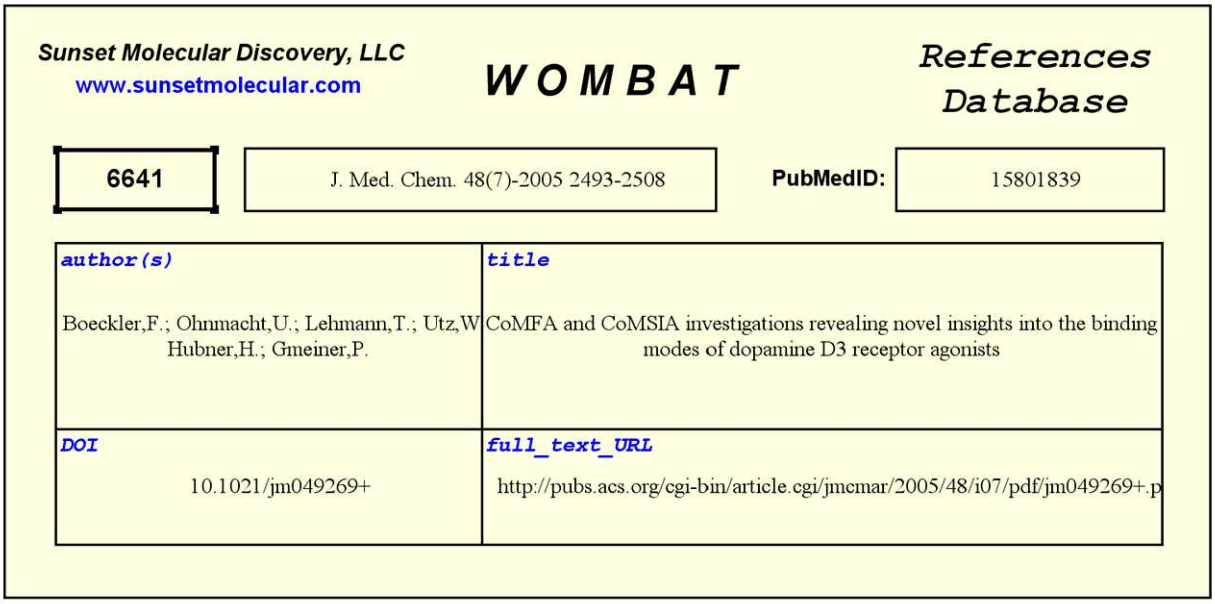


Figure 7. WOMBAT references database (example).

WOMBAT Database Structure

WOMBAT is a dynamic database, which evolves as new data types are included. The database structure is, however, preserved as much as possible from one release to the next. Each root record (or WOMBAT entry) is identified by a unique number (SMDLID), and is defined by the combination of *one chemical structure* and one or more associated *biological activities* as entered in *one publication* (Figure 3). One field, series ID (SID), links all the root records indexed from one reference (article). There are 6,801 SID values in WOMBAT 2006.1 (see also Figure 7). At the root level, information about the bibliographic reference (unique SID) from which the entry originated is recorded together with various properties (illustrated in Figure 6).

Separate keywords describe structural characteristics, related to stereochemistry (e.g., absolute, relative, +/-, R/S, 'non-chiral' or racemic) and to the salt form – see also Figure 3. We record the salt separately in order to avoid the salt-removal step that is usually performed in cheminformatic studies prior to structure computations. For each SMDLID, we define the following biological activity sub-records: AID, the activity identifier, with values from 1 to *n*, where *n* is the number of biological activity determinations for one structure; TargetName (the target name on which the activity was measured); ActType (the activity type, e.g., IC_{50}), ValueType, which can be one of five types: Exactly (=), lower than (<), greater than (>), % inhibition at a given concentration (@), or inactive;

ActValue, the numeric value of the activity, in negative $\log_{10}$ of the molar concentration; Range, the experimental confidence range for the measured activity, also in logarithmic units. For each SMDLID and each AID, we also record a number of BioKeywords related to biological activity information (e.g., bio-species, tissue and cell type, etc.) and target related information (e.g., the E.C. number [19], what radio-labeled substrate or ligand was used, etc.) – see also Figure 5. Thus, for one series (same SID value), each activity block (AID range 1, ..., *n*) has separate TargetName, ActType, ValueType and BioKeywords.

WOMBAT Quality Control

Quality control is performed at the moment of data entry, in particular with respect to errors present in publications. Chemical structures are checked for structural consistency by matching the molecular weight (MW) and chemical formula with the ones available in the Experimental section and/or Supporting Information – whenever available, and by comparison to prior publications. Whenever in doubt, we also use other sources, such as the Merck Index [25] and free Internet resources. In the instances where external and literature data cannot be reconciled, SciFinder [1] is also used. The error rate so far in medicinal chemistry publications is not at all negligible: We find an average of approximately 2 errors per publication in all the 6,791 papers indexed in WOMBAT 2006.1. Given the median of 25 compounds per series, this implies an overall error rate of 8%. These errors are distributed as follows [26]:

- incorrectly drawn or written structures (3%); incorrect molecular formula or molecular weight (3%);
- unspecified position of attachment of substituents, or ambiguous numbering scheme for the heterocyclic backbone (0.9%);
- structures with the incorrect backbone (0.7%);
- incorrect generic names or chemical names (0.2%);
- duplicates (0.2%);
- incorrect biological activity (0.3%);
- incorrect references (0.2%).

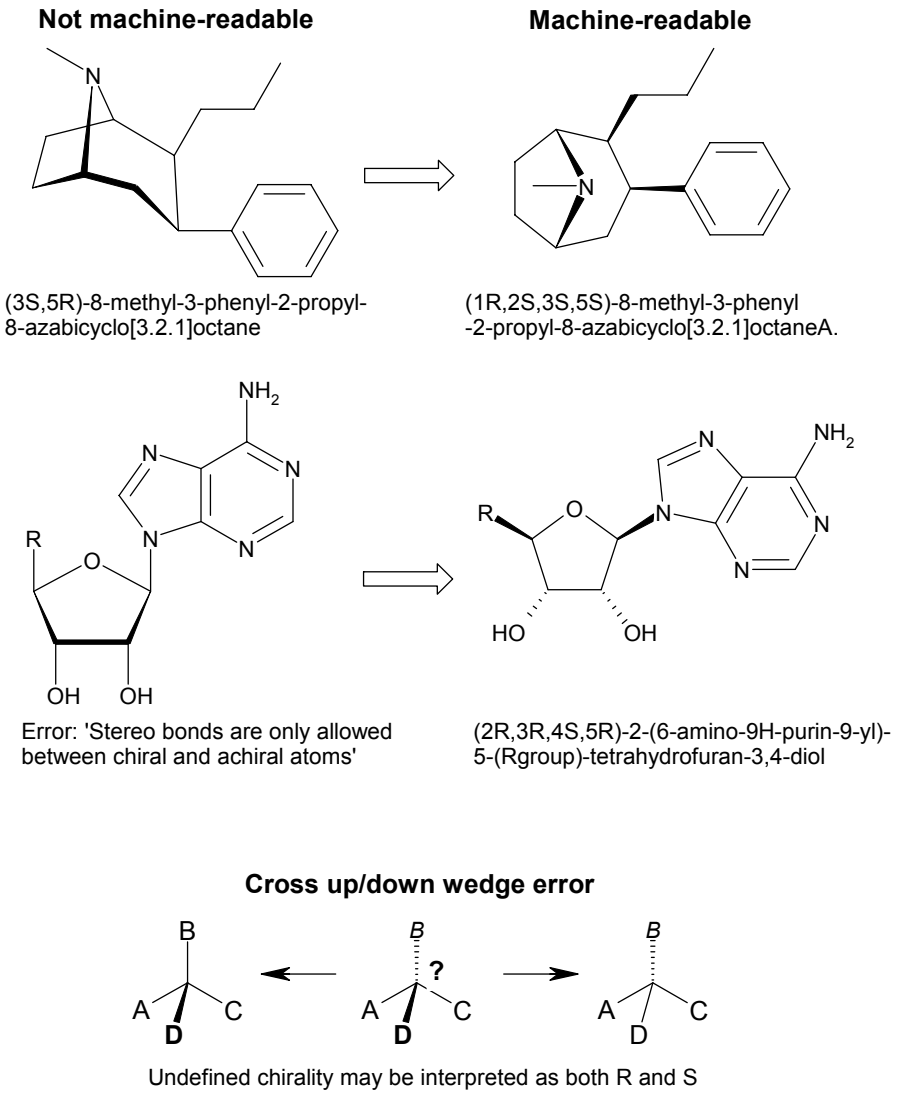


Figure 8. Human vs. machine-readable chemical structure representations. Names based on the depicted structures were interpreted using ACDName [27]. The cross up/down wedge error (middle) causes errors in assigning the absolute chirality.

Special attention is given to stereochemistry, as some compounds are published without proper chirality representation even though the information is available, e.g., for natural compounds and their derivatives. Furthermore, as illustrated in Figure 8, compounds published in medicinal chemistry literature are often depicted in a “human-readable” format; i.e., structures are drawn in a format that chemists can interpret to reconstruct proper chirality. However, this format is not “machine-readable”, i.e., cheminformatics software for 3D structural conversion, or for automatically generating IUPAC (International Union of Pure and Applied Chemistry) nomenclature, cannot perceive the stereo centers correctly if the “above/below plane” convention is not strictly enforced. We illustrate this with ACDName [27] on the structures depicted in Figure 6: The

software does not perceive two stereo centers for the tropane ring on the left side and returns an error for the sugar structure. The errors are not specific to ACDName – this program is used only to illustrate the problem. Another type of problem in structure-conversion is the cross up/down wedge error, when two such bonds emerge from the same chiral center (Figure 6): Software cannot assign the proper chirality, since by convention 3 atoms are in the 'paper plane', and only one is 'wedged' (up or down); two wedged bonds are simply not possible according to the convention. Most of these errors can be corrected by checking previous literature. Sometimes, even the cited reference may turn out to be in error, e.g., the reported MW is not consistent with the drawn, or named, structure.

From a quality control standpoint, the assignment of the SwissProt ID for each target can be a challenge, as publications do not always specify the exact target used in an assay. In some instances, the species from which the target was isolated is not explicitly mentioned, whereas some publications do not mention what target subtype was used. For example, there are 1,780 entries in WOMBAT 2006.1 that contain 'estrogen receptor' (ER) in the target name, which implies that ERs present in a particular organ (e.g., uterus, breast, brain) were tested for binding, agonism or antagonism. Of these, 1,201 entries were annotated for a specific receptor subtype, either ERα or ERβ, or '3A1' and '3A2' according to the nuclear receptors nomenclature [28]. For the remaining 579 entries, a target could not specifically be assigned to a single SwissProt ID.

This begs the question of storing multiple SwissProt ID values when a mixture of targets is present. This situation is common for, e.g., integrin receptors, which have the 2 protein chains separately defined in SwissProt. In the ER example, 114 of the 579 entries were tested on MCF7 cells; however, it is now clear that a third estrogen receptor, GPR30 [29], could be present in MCF7 cells [30]. Therefore, the observed anti-estrogenic activities for these 114 entries should be questioned in the light of this new information; should 3 such receptors be encoded? It further illustrates the dynamic nature of biological targets: As biologists uncover more information about a particular target or class of targets, and as our understanding about each target evolves, the exact nomenclature changes as well. For example, there are 852 entries in WOMBAT 2006.1 that contain 'VEGFR-2' as the target name: This target name stands for the vascular endothelial growth factor receptor subtype 2, but was previously known as 'Flk-1/KDR',

or ‘fetal liver kinase-1’ and ‘kinase insert domain-containing receptor’. The VEGFR-2 name is present in all 852 entries, even though some of the older (before 1999) publications did not refer to this target by the VEGFR-2 name. In an annotated database such as WOMBAT, one has to monitor and update not only changes related to biology, but also changes related to chemistry (and chemical errors), discussed in more detail below. Practical applications based on WOMBAT data mining using targets [31] and descriptors [32] have been described.

Uncovering Errors From Literature

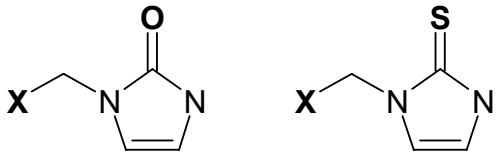
As the demand for integrated chemical and biological information increases, scientists rely more often on annotated databases that capture medicinal chemistry literature (see Tables 1 and 3). There is little, if any, error checking downstream from publication time, even though mechanisms for publishing errata have been in place for quite some time. While the responsibility for published data accuracy resides primarily with the author(s), it is also the responsibility of annotated database curators to capture as many of these errors as possible. While ensuring the quality control in WOMBAT, we have found inconsistencies in many of these publications. These errors may have a significant effect on the way we understand the molecular basis of chemical-biological interactions, at least for some particular series used for structure-activity studies. Coats has traced the errors on a known steroid benchmark for QSAR studies to the original publications [33]. Some of these errors are discussed below.

Example 1. The following errors were found in Table 1 of reference [34], page 126:

- compounds with molecular names **53** and **56**, respectively, appear to be duplicated because all their substituents are identical. Based on their activities, **56** (compound **15e** in [35]) has the *meta* $–OCH_3$-C_6H_4 substituent, while **53** (compound **15g** in [35]) has the ***para*** $–OCH_3$-C_6H_4 substituent;
- the –NH– group is missing from the L substituent in compound **27** (compound **9** in [36]), and the $–CH_2–$ group is missing from the L substituent in compound **45** (compound **13** in [35]);
- the R substituent of compound **66** is $–C_6H_2$-2-CO_2CH_3-4,5-$(CH_3)_2$ instead of the correct $–C_6H_2$-2-CO_2CH_3-4,5-$(\mathbf{O}CH_3)_2$ group (compound **5b** in [36]);

- the R substituent of compound **68** is $–C_6H_2$-2-CO_2-4,5-$(CH_3)_2$ instead of the correct $–C_6H_2$-2-CO_2-4,5-(**O**$CH_3)_2$ group (compound **7** in [35]);
- compound **44** has a $–\log(IC_{50})$ of 7.67 instead of the correct 7.74 (compound **15d** in [35]).

Example 2. In Table 1b of reference [37], page 4361, the core structure contains an oxygen-atom instead of the correct sulfur-atom [38]:



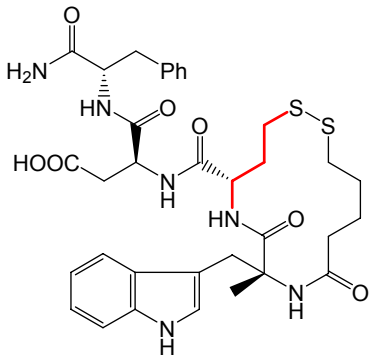
wrong correct

Thus, 47 structures (where X is the rest of the molecule) are incorrect in [37]. Since the paper illustrates the capabilities of a particular structure-activity method, the consistent error does not influence the validity of the models; it would, however, greatly influence the use of this series/model in a medicinal chemistry project where the goal would be to improve the binding affinity. Starting from the same initial publication [38], other errors were propagated in [39]:

- compound **37** has an incorrect double substitution in the para position of the aromatic ring, 2,4-NO_2,4-OH, while the correct one is 3-NO_2,4-OH;
- the R substituent of compound **B.12** is 2,4,6-Cl_2,4-OMe instead of the correct 2,6-Cl_2,4-OMe.

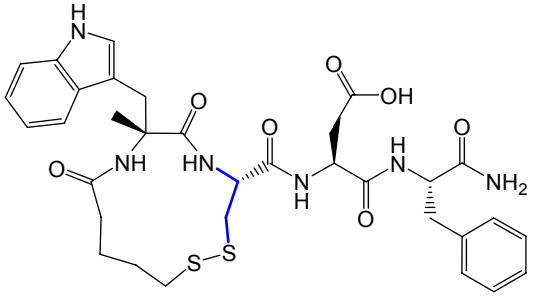
Example 3. Errors could be found also in Chemical Abstracts' SciFinder [1]. All the errors we encountered originate in the primary publications; their appearance in SciFinder illustrates how such errors can propagate (since SciFinder is a very popular resource). For example, the compound **RB-380** (CAS# 187454-94-0), published in [40] (original molecule name **24**) has a ring size of 14 atoms, instead of the correct 13:

SciFinder structure Correct structure



$C_{34}H_{42}N_6O_7S_2$

L-Phenylalaninamide, N-(5-mercapto-1-oxopentyl)-α-methyl-D-tryptophyl-L-homocysteinyl-L-α-aspartyl-, cyclic (1→2)-disulfide (this name is given in CAS)

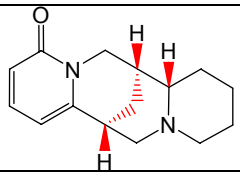
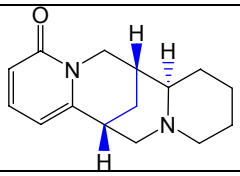
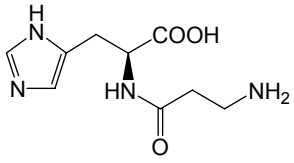
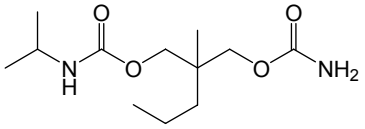


$C_{33}H_{40}N_6O_7S_2$

Cyclo-*S,S*-[(5-thiopentanoyl)-αMe(R)-Trp-Cys]-Asp-Phe-NH_2 (this name is given in the original publication [40], in the experimental section)

The correction we propose is based on the experimental section name and on the following text fragment (p.648, results section [40]): *"... by introducing an additional amide bond (compound **16** or RB 370) or a disulfide bridge (**24** or RB 380) into the 13-membered ring (Schemes 2 and 3), and by changing the size of the ring (Table 1, compounds **43** and **45**)."* By analyzing the data from Table 1 [40], compound **43** (which is actually **44** – which is another small error) has a 13 atoms ring, while compound **45** has a 14 atoms ring.

Example 4. Stereochemical ambiguities and structural errors can be encountered in the Merck Index [25] as well, as shown in these two examples:

Compound identifiers & error description	Merck Index structure	Correct structure
M630, anagyrine (CAS# 486-89-5): chiral center inversion & cross up/down wedge		
M1854, carisoprodol (CAS# 78-44-4): completely different structure. All other information about M1854 is correct (name, formula and molecular weight). The formula is correct in the 9th edition of the Merck Index.		

The examples from SciFinder and the Merck Index are not intended to question the quality of these products, which we consider to be outstanding. They are invaluable

resources to many chemists worldwide, and the error rate in these two databases is insignificant if one takes into account the enormous volume of indexed data. We have published a structure-activity paper on HIV-protease inhibitors [41] where a modified peptide was present in both the training set, and the test set. Al Leo of Pomona College has recently [42] detected 100 chemical and name errors in the printed version of the sixth edition of Burger's Medicinal Chemistry [43], errors that are be corrected in the on-line edition [44]. One can never be too careful in verifying the available information, in particular if one is to invest a significant amount of resources in that area.

WOMBAT-PK: Clinical pharmacokinetics (PK) and toxicological (Tox) data

As PK data become more important during lead discovery and evaluation, we screened the clinical pharmacokinetics literature and developed a chemical database that captures such data in numerical searchable format (WOMBAT-PK). Its organization is illustrated in Figures 9-11, which illustrate three of the 4 panels of the database: The *Compound Description* panel (Figure 9) provides the drug marketed names, some physico-chemical characteristics, as well as structural chemical (2D depiction and SMILES code) and related information (chirality, salt). The *Pharmacokinetic Data* panel (Figure 10) provides the drug target information, and multiple PK and Tox parameters, indexed in both numerical and text form. The third panel, *Potential Side Effects*, captures data for BBB (blood-brain barrier) permeability, cardiac toxicity data, possibly related to hERG (human ether-a-go-go potassium channel 1) bioactivity, in vitro bioactivities from WOMBAT, as well as mammalian tox data (e.g., the lethal dose 50%, LD_{50}). The fourth panel, *Computed Chemical Properties* panel, is identical to the one in WOMBAT (see Figure 6). The 2006 release of WOMBAT-PK contains 900 marketed drugs (in rare cases, some are metabolites) with documented PK and Tox properties.

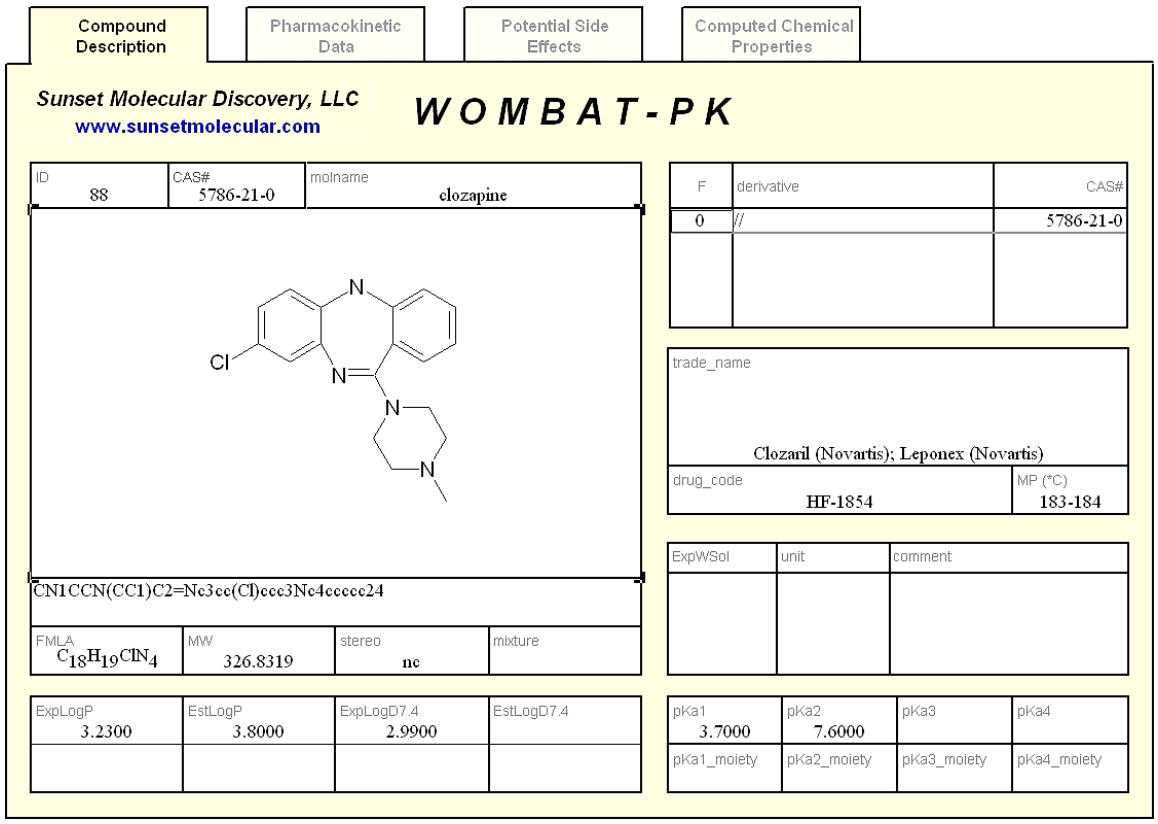


Figure 9. WOMBAT-PK compound description panel (example).

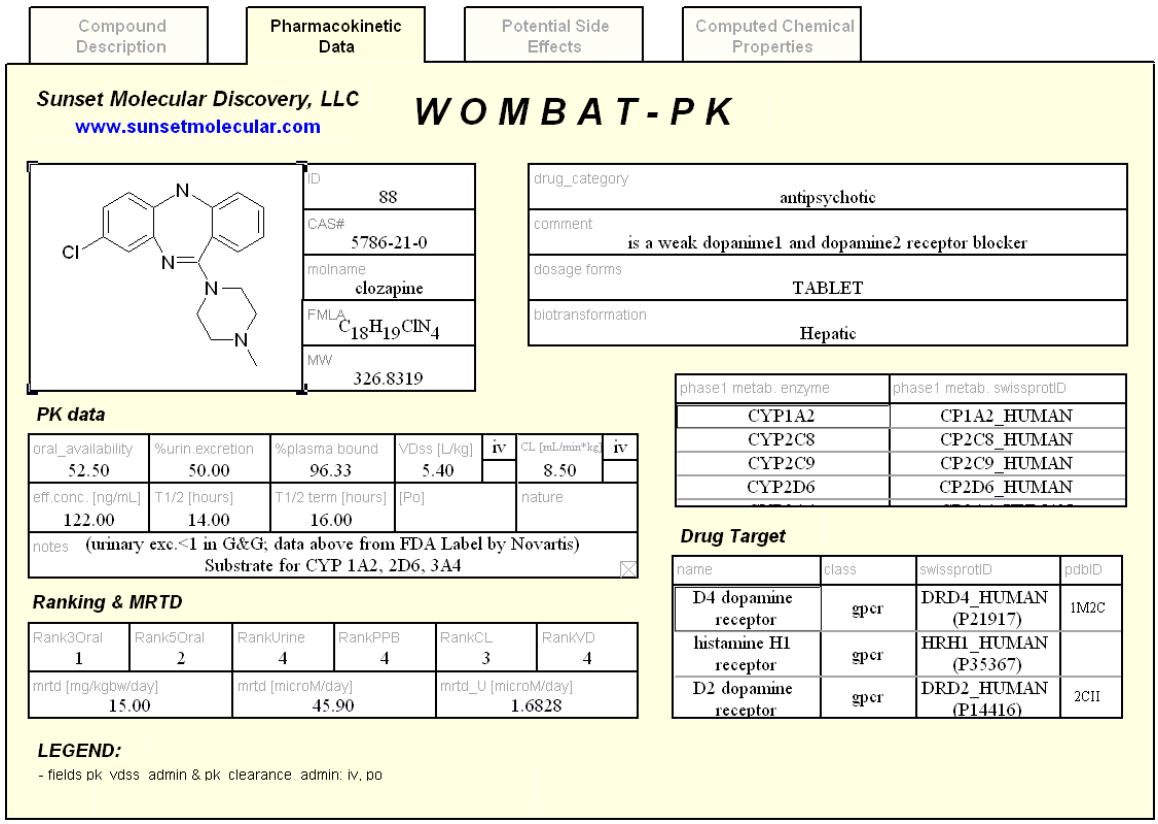


Figure 10. WOMBAT-PK pharmacokinetic data panel (example).

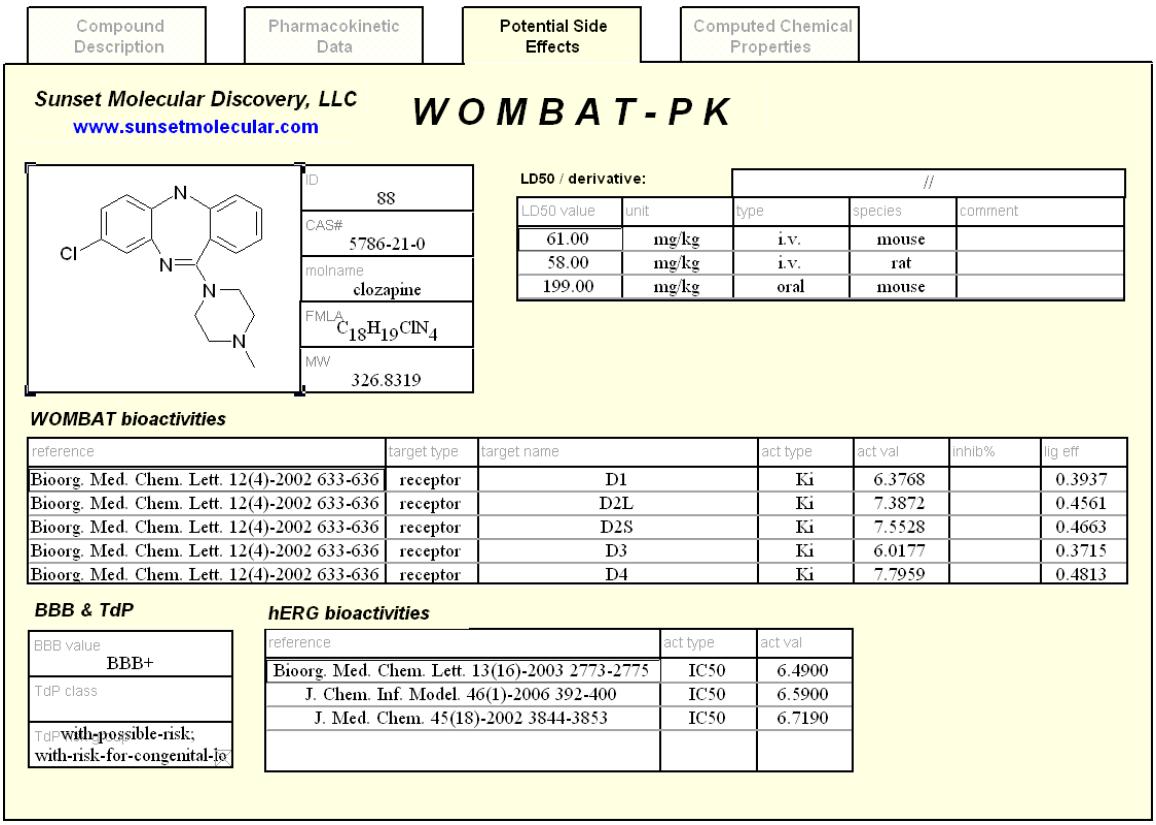

Compound Description
Pharmacokinetic Data
Potential Side Effects
Computed Chemical Properties
Sunset Molecular Discovery, LLC
www.sunsetmolecular.com
W O M B A T - P K
ID 88
CAS# 5786-21-0
molname clozapine
FMLA $C_{18}H_{19}ClN_4$
MW 326.8319
LD50 / derivative: //
LD50 value | unit | type | species | comment
61.00 | mg/kg | i.v. | mouse
58.00 | mg/kg | i.v. | rat
199.00 | mg/kg | oral | mouse
WOMBAT bioactivities
reference | target type | target name | act type | act val | inhib% | lig eff
Bioorg. Med. Chem. Lett. 12(4)-2002 633-636 | receptor | D1 | Ki | 6.3768 | | 0.3937
Bioorg. Med. Chem. Lett. 12(4)-2002 633-636 | receptor | D2L | Ki | 7.3872 | | 0.4561
Bioorg. Med. Chem. Lett. 12(4)-2002 633-636 | receptor | D2S | Ki | 7.5528 | | 0.4663
Bioorg. Med. Chem. Lett. 12(4)-2002 633-636 | receptor | D3 | Ki | 6.0177 | | 0.3715
Bioorg. Med. Chem. Lett. 12(4)-2002 633-636 | receptor | D4 | Ki | 7.7959 | | 0.4813
BBB & TdP
BBB value BBB+
TdP class
TdP with-possible-risk; with-risk-for-congenital-lo
hERG bioactivities
reference | act type | act val
Bioorg. Med. Chem. Lett. 13(16)-2003 2773-2775 | IC50 | 6.4900
J. Chem. Inf. Model. 46(1)-2006 392-400 | IC50 | 6.5900
J. Med. Chem. 45(18)-2002 3844-3853 | IC50 | 6.7190

Figure 11. WOMBAT-PK potential side effects panel (example).

Table 4. Experimental PK and Tox data captured in WOMBAT-PK 2006.1.

	On File	G & G	Avery
%Oral Bioavailability	740	312	277
%Urinary excretion	339	326	N/A
%Plasma Protein Binding	776	311	434
Clearance, Cl (mL/min*kg)	514	320	422
Nonrenal clearance (fractional)	442	N/A	442
Volume of distribution, VDss (L/kg)	552	322	453
Half-life, T1/2 (hrs)	839	338	576
Terminal half-life, TT1/2 (hrs)	581	N/A	580
Effective concentration (mM/L)	119	118	N/A
MRTD (µmole/kg-bw/day)	575	N/A	N/A
MRTD_U (µmole/kg-bw/day; f_u corrected)	498	N/A	N/A
LogD7.4 (measured)	513	N/A	N/A
LogPoct (measured)	472	N/A	N/A
pKa1	350	N/A	274
pKa2	99	N/A	75
In Vitro Binding Data (from WOMBAT)	453	N/A	N/A

Currently indexed PK, Tox and physico-chemical properties data are summarized in Table 4. The top 9 properties were captured from the following sources: Goodman & Gilman's 9th edition [45] (G&G), Avery's 4th edition [46] (Av), and the Physician Desk Reference [11] (PDR). FDA's Center for Drug Evaluation and Research website [47]

was consulted for FDA approved drug labels. Other resources (e.g., Google™) were sometimes used to compile the WOMBAT-PK database. The maximum recommended therapeutic dose [48] (MRTD) is available from the FDA [49], whereas MRTD_U (MRTD corrected for the fraction-unbound) was determined by using the %PPB data already indexed in WOMBAT-PK. Thus, MRTD_U = MRTD * (1 - %PPB), and is available for 498 drugs. Experimental $LogD_{7.4}$ and LogP values from compilation tables [50] and from the Sangster database [51], and pK_a values from Avery [46] and the Merck Index [25] were collected for these drugs. In WOMBAT-PK, drug targets are assigned to 753 drugs (of these, 97% have SwissProt IDs), whereas the phase I metabolizing enzymes (all with SwissProt IDs) are recorded for 419 entries. Regarding cardiac toxicity, there are 218 drugs indexed for QT-prolongation (a clinical observation based on the ECG, the electrocardiogram), 89 for *Torsade de Pointes* risk (another ECG signal), and 71 with hERG binding data.

Table 5. Parent value ranking for certain PK parameters in WOMBAT-PK 2006.1.

%Oral	Rank3Oral	Rank5Oral
0-5	0	0
5.1-19.99	0	1
20.0-79.99	1	2
80.0-95	2	3
>95.1	2	4

%Urine	RankUrine
0-1	0
1.01-5	1
5.01-20	2
20.01-50	3
50.01-80	4
>80	5

%PPB	RankPPB	Cl (mL/min*kg)	RankCl
0-5	0	0-(6/7)	0
5.01-20	1	(6.01/7)-(12/7)	1
20.01-80	2	(12.01/7)-5	2
80.01-95	3	5.01-10	3
95.01-99	4	10.01-15.5	4
>99.1	5	>15.5	5

VD (L/kg)	RankVD
0-1	0
1.01-3	1
3.01-5.5	2
5.51-12	3
>12	4

Curating clinical PK data requires individual examination [52], and sources such as Goodman-Gilman are often considered more reliable. Often, such experimental values are “greater than” or “less than” a given cut-off value. A systematic round-off procedure was implemented, whereby “<5” was attributed a higher value (=2.5), compared to “<1” (=0.5). Numeric values also differ, sometimes significantly, due to various factors (e.g., multiple dose vs. single-dose, children vs. healthy volunteers, etc.), thus conflicting values were sometimes reported. The “on file” values in Table 4 are often averages between G&G and Av data, although ≈30% of the indexed values differ by more than 20% between these two sources (data not shown). To identify trends we attenuated the

effect of such discrepancies by implementing an incremental increase procedure to some of the PK properties, as illustrated in Table 5. Incremental rank values were selected from experience whenever possible: e.g., experimental errors related to %Oral occur mostly for values between 20 and 80%; 6/7 and 12/7 represent the ½ and full value of creatinine clearance (120 mL/70 kg/min), respectively; 3, 5.5 and 12 are typical 70-kg man volumes in liters for plasma, blood and extracellular fluids [14].

WOMBAT-PK also captures information about the known (or intended) drug target(s). These are often retrieved from the therapeutic classification data (e.g., anti-histaminic compounds are intended to act as antagonists of the H1 histamine receptor), or can be inferred by searching medicinal chemistry literature – see also Figure 10. Of interest is the cross-index of the WOMBAT and WOMBAT-PK databases, which shows *in vitro* binding information for certain drugs in medicinal chemistry literature. For example, aspirin has a relatively weak binding affinity to cyclooxygenases COX-1 and COX-2 (but acts as suicide inhibitor); in the same time, it appears to be 2-3 orders of magnitude more potent on GP IIb/IIIa, an $\alpha_{2b}\beta_{3a}$ integrin involved in platelet aggregation. This probably explains why aspirin is effective at the 75-80 mg/day dose range as an antiaggregant, compared to the 500-1000 mg/day dose range for the anti-inflammatory effects [53].

Datamining With WOMBAT

Table 6. Distribution of target types among oncology targets in WOMBAT 2006.1.

Target Type	Count	Percent
enzyme	759	82.77
ion channels	4	0.44
protein	56	6.11
receptor	98	10.69

Example 1. One of the major areas of interest in medicinal chemistry is oncology. The cancer medicinal chemistry space was described earlier by mining the WOMBAT and WOMBAT-PK databases [54]. The *oncology* subset of WOMBAT 2006.1 contains active on 917 unique targets, detailed in Table 6. A query for targets that have over 300 entries allows us to establish an activity histogram, contrasting low-activity entries (Figure 12a) with high-activity entries (Figure 12b). This allows the user to rapidly identify targets for which the number of low-activity entries exceeds significantly the number of to high-activity entries, e.g., GGTase, PKA and Tubulin - see Figure 12 legend for target names.

In fact, there are only 7 entries of Tubulin inhibitors with activity better than 100 nM, and only 2 of them are Ro5 [21] compliant (data not shown). One can conclude that such targets are areas of opportunity for the design of novel inhibitors. By the same token, AR, ERβ and MMP-13, respectively, are targets where the number of high-activity entries highly exceeds the low-activity records. These targets are, probably, already abundant with high-quality ligands, indicating that perhaps selectivity or pharmacokinetic profiling are currently the key areas for further optimization.

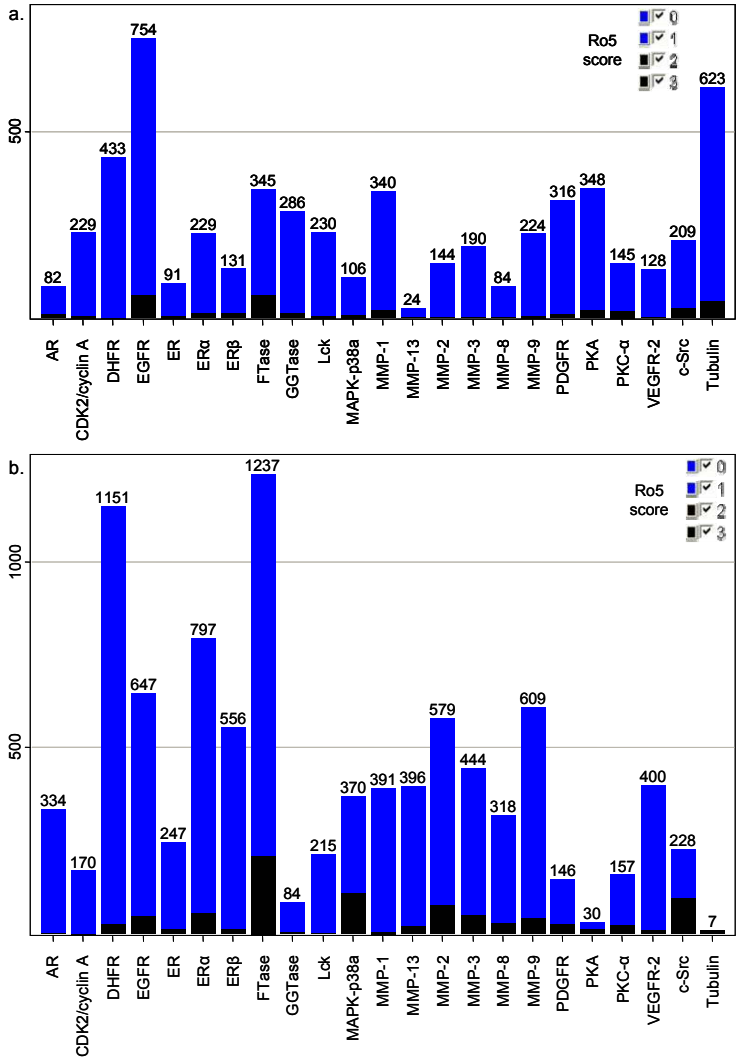


Figure 12. Activity histogram for the most-populated oncology-related targets in WOMBAT 2006.1. There were at least 300 records per target. The top panel (a.) shows low-activity compounds (10 μM or less), whereas the bottom panel (b.) shows high-activity compounds (100 nM or better). The bars are color-coded according to Ro5 violations (see also legend). Numbers on top of each bar indicate the number of componds with low (a.) and high (b.) entries per target. Target names are as follows: AR: androgen (or dihydrotestosterone) receptor; CDK2/cyclin A: cell division protein kinase 2; DHFR: dihydrofolate reductase; EGFR: epidermal growth factor receptor; ER, ERα and ERβ: Estrogen receptor, ER alpha and beta subtypes, respectively; Ftase: protein farnesyltransferase; GGTase: protein geranyl-geranyltransferase; Lck: proto-oncogene tyrosine-protein kinase Lck; MAPK p38: cytokine suppressive anti-inflammatory drug binding protein, or mitogen-activated protein kinase p38 alpha; MMP-1 through MMP-9: matrix

metalloproteases 1 through 9, respectively; PDGFR: platelet derived growth factor receptor; PKA: cAMP-dependent protein kinase A; PKC-alpha: protein kinase C, alpha type; VEGFR-2; vascular endothelial growth factor receptor 2, or kinase insert domain receptor; c-Src: proto-oncogene tyrosine-protein kinase SRC; and Tubulin.

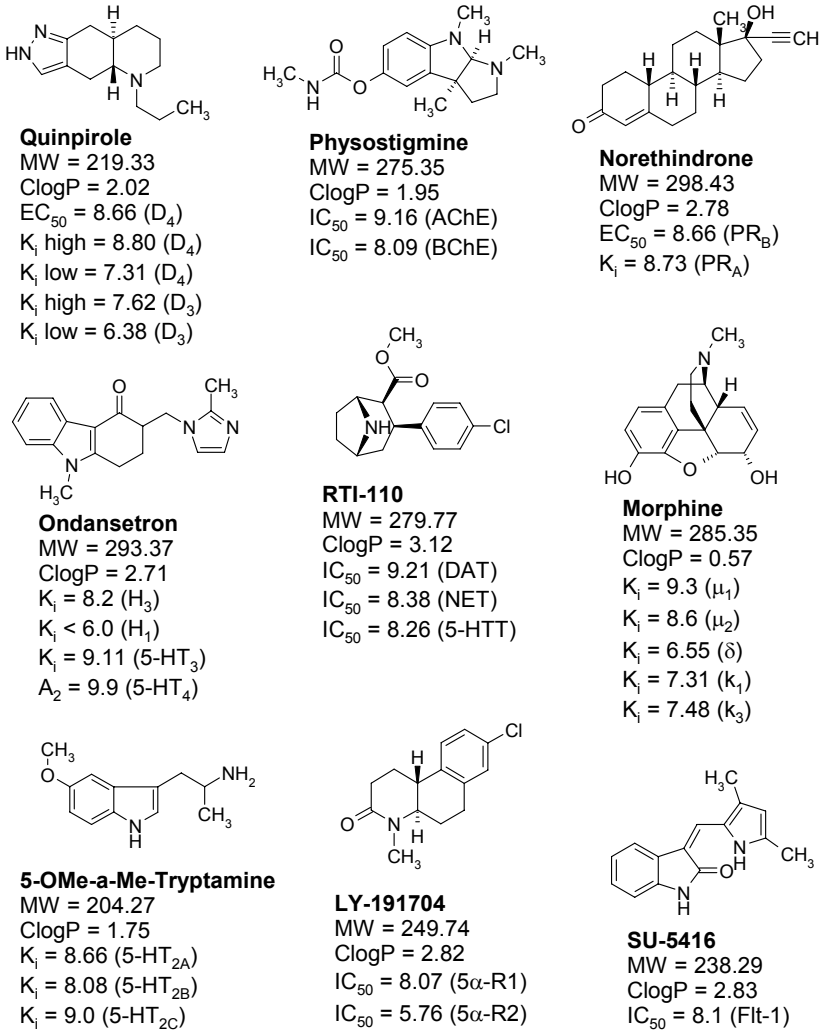


Figure 13. Examples of rule-of-three compliant molecules that have biological activity better than 10 nM. Under each molecule, the following information is included: molecule name, MW, ClogP, the biological activity type, value and target. Target names are as follows: D_3 and D_4 – dopaminergic receptor types 2 and 3; AChE and BChE – acetyl- and butyryl- choline esterases; PR_A and PR_B – progesterone receptor types A and B; H_1 and H_3, histamine receptor types 1 and 3; 5-HT_{2A}, 5-HT_{2B}, 5-HT_{2C}, 5-HT_3, 5-HT_4 – serotonin receptor subtypes 2A, 2B, 2C, and types 3 and 4; DAT, NET, 5-HTT – dopamine, norepinephrine and serotonin transporter proteins; μ_1, μ_2, δ, k_1, k_3 – opioid receptor types mu-1, mu-2, delta, kappa-1 and kappa-3; 5α -R1 and 5α-R2 – 5-alpha-reductase isozymes 1 and 2; Flt-1 - fms-like tyrosine kinase receptor.

Example 2. The concept of leadlikeness [32, 55, 56] and its application in developing leadlike libraries [55, 57, 58] have been extensively discussed. The reduction of the leadlike concept into practice at Astex [59] resulted in a proposal for fragment libraries in lead discovery called the 'Rule of Three': MW < 300, ClogP < 3, number of hydrogen bond donors and acceptors less than, or equal to 3, three or less flexible bonds, and

PSA ≤ 60 $Å^2$. Using these criteria, WOMBAT 2006.1 returns 6,607 entries. Of these, 2,001 entries contain at least one biological activity better than, or equal to 100 nM, and 543 of these contain a generic name. This usually means they are either launched drugs, or natural products, or otherwise in an advanced stage of development. The examples given in Figure 13 illustrate the chemotype, target and activity diversity that can be found in rule-of-three compliant molecules: Neurotransmitter and nuclear hormone receptor agonists (EC_{50}) and antagonists (K_i, IC_{50} and A_2), neurotransmitter transporters, as well as enzyme inhibitors are present, most of them with multiple activities. Based on the WOMBAT 2006.1 entries, it appears that there are a number of interesting chemotypes that are rule-of-three compliant. Such cheminformatics-based mining can identify target-specific small molecules for fragment library design [60].

Conclusions and Future Challenges

As annotated databases, WOMBAT and WOMBAT-PK continue to evolve in time – not only with the addition of more entries, but also with updates and restructuring of the biological, clinical and chemical information, which is subject to revision even after the data are captured and indexed. The inclusion of the pre-computed properties panel allows the users to quickly identify rule-of-five or rule-of-three compliant datasets, or to constrain the query with respect to, e.g., flexible bonds, polar surface area, computed solubility or LogP, etc. WOMBAT and WOMBAT-PK are currently available in the MDL Isis/Base format. WOMBAT is also integrated in CABINET (Chemical And Biological Informatics NETwork) [61, 62] as a server. CABINET [62], a federation of high-performance scientific databases which collaborate via web-like interfaces in order to provide integrated access to diverse chemical and biological information, was described elsewhere [61].

Federated database servers such as CABINET could, for example, bring together WOMBAT and C-QSAR [63], but the challenge goes beyond technical issues related to field correspondence. Data normalization (e.g., ensuring similar treatment regarding chirality, salt information, measured and computed properties) is likely to require on-the-fly data interpreters, which in turn forces lack of ambiguity for all data entries in WOMBAT and other databases. Data transparency is not always possible: For example, most WOMBAT entries related to EGFR (epithelial growth factor receptor) are classified

as ‘TargetType = enzyme’, because EGFR is a membrane receptor-linked tyrosine-protein kinase and medicinal chemists target EGFR for kinase inhibition. However, in one instance, ‘TargetName’ was assigned as ‘receptor’ because the endogenous ligand, EGF, was used to test for EGFR antagonism [64]. Thus, restricting data fields to certain value types, usually an asset for database indexing, can become a hindrance when the unexpected occurs. Thus, one of the challenges in database federation remains adaptive data normalization for biology-related data fields, since biological phenomena are not always amenable to unambiguous mapping. By successfully addressing these problems, it is quite likely that integrated data mining tools will change the way we conduct everyday research.

Acknowledgments

The authors thank Prof. Hugo Kubinyi (Heidelberg, Germany) for suggestions.

Reference

1 Chemical Abstracts online and its search module, SciFinder, are available from the American Chemical Society, http://www.cas.org/SCIFINDER/, 2006
2 The Beilstein Information Systems is available from http://www.beilstein.com/, 2006
3 The Spresi Database is available from InfoChem GmbH, München, http://www.spresiweb.de/; and from Daylight Chemical Information Systems, http://www.daylight.com/products/databases/Spresi.html, 2006
4 MDDR is available from MDL Information Systems, http://www.mdli.com/products/finders/database_finder/ and from Prous Science Publishers, http://www.prous.com/index.html, 2006
5 WDI, the Derwent World Drug Index, is available from Derwent Publications Ltd., http://thomsonderwent.com/products/lr/wdi/ and from Daylight Chemical Information Systems, http://www.daylight.com/products/databases/WDI.html, 2006
6 The Current Patents Fast Alert database is available from Current Patents Ltd., London, http://www.current-patents.com/, 2006
7 The Comprehensive Medicinal Chemistry database is available from MDL Information Systems, Inc., http://www.mdli.com/products/knowledge/medicinal_chem/index.jsp, 2006
8 DiscoveryGate is available from MDL Information Systems, Inc., http://www.mdli.com/products/knowledge/discoverygate/; a subset of DiscoveryGate is available through the PubChem system, see http://www.mdli.com/company/news/press_releases/2006/pr_pubchem_21mar06.jsp , 2006
9 The PubChem database is available online at the National Center for Biotechnology Information, http://pubchem.ncbi.nlm.nih.gov/, 2006
10 Austin, C.P., Brady, L.S., Insel, T.R., and Collins, F.S. NIH molecular libraries initiative. *Science* **2004**, *306*, 1138-1139.
11 The Physician Desk Reference is produced by Thomson Healthcare, 2003, ISBN 1-56363-472-4, and is available online at http://www.pdr.net/, 2006
12 The DrugBank database is available at http://redpoll.pharmacy.ualberta.ca/drugbank/ , 2006

13 Olah, M., Mracec, M., Ostopovici, L., Rad, R., Bora, A., Hadaruga, N., Olah, I., Banda, M., Simon, Z., Mracec, M., Oprea, T.I. WOMBAT: World of Molecular Bioactivity. In Cheminformatics in Drug Discovery, Oprea, T.I. (Ed.). Wiley-VCH, New York, **2005**, 223-239
14 Oprea, T.I. Benedetti, P., Berellini, G., Olah, M., Fejgin K., Boyer, S. Rapid ADME Filters for Lead Discovery. In Molecular Interaction Fields, Cruciani, G. (Ed.). Wiley-VCH, New York, **2006**, 249-272
15 WOMBAT and WOMBAT-PK are available from Sunset Molecular Discovery, Santa Fe, New Mexico, http://www.sunsetmolecular.com , 2006
16 Olah M., Oprea, T.I. Bioactivity Databases. In Comprehensive Medicinal Chemistry II, Taylor, J., Triggle, D. (Eds.). Elsevier, New York, **2006**, 000-000
17 Weininger, D. SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36
18 Weininger, D., Weininger, A. and Weininger, J.L. SMILES 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97-101
19 The Enzyme Nomenclature is recommended by the International Union of Biochemistry and Molecular Biology, and is available at http://www.chem.qmul.ac.uk/iubmb/enzyme/, 2006
20 Swiss-Prot Protein knowledgebase database, http://kr.expasy.org/sprot/, 2006
21 Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Reviews* **1997**, *23*, 3-25
22 Leo, A. Estimating $LogP_{oct}$ from structures. *Chem. Rev.* **1993**, *5*, 1281-1306
23 Tetko, I. V. and Tanchuk, V. Y. (2002) Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. J. Chem. Inf. Comput. Sci. 42, 1136-1145; http://146.107.217.178/lab/alogps/index.html
24 The Digital Object Identifier (DOI) is a system for identifying and exchanging intellectual property in the digital environment (http://www.doi.org/). An object is directly accessible using the customized address http://dx.doi.org/DOI_VALUE, 2006
25 Merck Index 13th edition, Merck & Co. Inc., Rahway NJ, **2001**.
26 Oprea, T.I., Olah, M., Ostopovici, L., Rad, R. and Mracec, M. On the propagation of errors in the QSAR literature In EuroQSAR 2002 - Designing drugs and crop protectants: Processes, problems and solutions. Ford, M., Livingstone, D., Dearden, J. and Van de Waterbeemd H. (Eds.), New York, Blackwell Publishing, **2003**, 314-315.
27 ACDName is available from Advanced Chemistry Development Inc., Toronto, Ontario, http://www.acdlabs.com
28 G-protein coupled receptors are classified according to the GPCR nomenclature available at http://www.gpcr.org/7tm, whereas nuclear receptors are annotated based on the NR nomenclature available at http://www.receptors.org/NR, 2006
29 Filardo, E. J., Quinn, J. A., Bland, K. I. & Frackelton, A. R., Jr. Estrogen-induced activation of Erk-1 and Erk-2 requires the G protein-coupled receptor homolog, GPR30, and occurs via trans-activation of the epidermal growth factor receptor through release of HB-EGF. *Mol. Endocrinol.* **2000**, *14*, 1649-1660
30 Revankar, C. M., Cimino, D. F., Sklar, L. A., Arterburn, J. B. & Prossnitz, E. R. A transmembrane intracellular estrogen receptor mediates rapid cell signaling. *Science* **2005**, *307*, 1625-1630
31 Nidhi, Glick, M., Davies, J.W., Jenkins, J.L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Modeling* **2006**, *46*, 000-000.
32 Oprea, T.I. Cheminformatics in lead discovery. In Cheminformatics in Drug Discovery, Oprea, T.I. (Ed.). Wiley-VCH, New York, **2005**, 27-42
33 Coats, E.A. The CoMFA steroids as a benchmark dataset for development of 3D-QSAR methods. In 3D QSAR in Drug Design. Volume 3. Recent Advances. . Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) Dordrecht: Kluwer/ESCOM, **1998**, 199-213

34 Chen, Q., Wu, C., Maxwell, D., Krudy, G.A., Dixon, R.A.F. and You, T.J. A 3D-QSAR analysis of in vitro binding affinity and selectivity of 3-izoxazolylsulfonylaminothiophenes as endothelin receptor antagonists. *Quant. Struct.-Act. Relat.* **1999**, *18*, 124-133.
35 Wu, C., Chan, M.F., Stavros, F., Raju, B., Okun, I., Mong, S., Keller, K.M., Brock, T., Kogan T.P., Dixon, R.A. Discovery of TBC11251, a potent, long acting, orally active endothelin receptor-A selective antagonist. *J. Med. Chem.* **1997**, *40*, 1690-1697
36 Wu, C., Chan, M.F., Stavros, F., Raju, B., Okun, I. and Castillo, R.S. Structure-activity relationships of N2-aryl-3-(isoxazolylsulfamoyl)-2-thiophenecarboxamides as selective endothelin receptor-A antagonists. *J. Med. Chem.* **1997**, *40*, 1682-1689
37 So, S.S. and Karplus, M. Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications. *J.Med. Chem.* **1997**, *40*, 4360-4371
38 Burke, B.J. and Hopfinger, A.J. 1-(Substituted-benzyl)imidazole-2(3H)thione inhibitors of dopamine β-hydroxylase. *J. Med. Chem.* **1990**, *33*, 274-281
39 Vedani, A. McMasters, D.R. and Dobler, M. Multi-conformational ligand representation in 4D-QSAR: Reducing the bias associated with ligand alignment. *Quant. Struct.-Act. Relat.* **2000**, *19*, 149-161
40 Blommaert, A.G.S., Dhotel, H., Ducos, B., Durieux, C., Goudreau, N., Bado, A., Garbay, C. and Roques, B.P. Structure-based design of new constrained cyclic agonists of the cholecystokinin CCK-B receptor. *J. Med. Chem.* **1997**, *40*, 647-658
41 Oprea, T.I., Waller, C.L. and Marshall, G.R. 3D-QSAR of human immunodeficiency virus (I) protease inhibitors. II. Predictive power using limited exploration of alternate binding modes. *J. Med. Chem.* **1994**, *37*, 2206-2215
42 Leo, A. Personal communication, 2004
43 Abraham, D. Burger's Medicinal Chemistry, 6th Edition. Wiley-VCH, New York, 2003
44 Abraham, D. Personal communication, 2004
45 Hardman, J.G., Limbird, L.E., Molinoff, P.B., Ruddon, R.W., Gilman, A.G. *Goodman & Gilman's the Pharmacological Basis of Therapeutics, 9th edition*. McGraw Hill, New York, **1996**.
46 Speight, T.M. and Holford, N.H.G. *Avery's Drug Treatment*, *4th edition*. Adis International, Auckland, **1997**.
47 FDA labels are at the Center for Drug Evaluation and Research (CDER) website, http://www.accessdata.fda.gov/scripts/cder/drugsatfda/, 2006
48 Contrera, J.F., Matthews, E.J., Kruhlak, N.L., and Benz, R.D. Estimating the safe starting dose in phase I clinical trials and no observed effect level based on QSAR modeling of the human maximum recommended daily dose. *Regulatory Toxicol. and Pharmacol.* **2004**, *40*, 185-206.
49 MRTD is available from the CDER website, http://www.fda.gov/cder/Offices/OPS_IO/MRTD.htm, 2006
50 Hansch, C., Leo, A., and Hoekman, D. *Exploring QSAR, vol. 2*, ACS Publishers, Washington DC, **1995**.
51 The Sangster database is available at http://logkow.cisti.nrc.ca/
52 Benet, L.Z. Personal communication, 2006
53 Andrieu, S., Lebret, M., Maclouf, J., Beverelli, F., Giudicelli, J.F., Berdeaux, A. Effects of antiaggregant and antiinflammatory doses of aspirin on coronary hemodynamics and myocardial reactive hyperemia in conscious dogs. *J. Cardiovasc. Pharmacol.* **1999**, *33*, 264-272
54 Lloyd, D.G., Golfis, G., Knox, A.J.S., Fayne, D., Meegan, M.J., Oprea, T.I. Oncology exploration: charting cancer medicinal chemistry space. *Drug Discov. Today* **2006**, *11*, 149-159
55 Hann, M.M. Leach, A. and Green, D.V.S. Computational chemistry, molecular complexity and screening set design. In Cheminformatics in Drug Discovery, Oprea, T.I. (Ed.). Wiley-VCH, New York, **2005**, 43-57
56 Hann, M.M. and Oprea, T.I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255-263.

57 Goodnow Jr., R.A., Gillespie, P. and Bleicher, K. Cheminformatic tools for library design and the hit-to-lead process: A user's perspective. In Cheminformatics in Drug Discovery, Oprea, T.I. (Ed.). Wiley-VCH, New York, **2005**, 381-435
58 Baringhaus, K.H. and Matter, H. Efficient strategies for lead optimization by simultaneously addressing affinity, selectivity and pharmacokinetic parameters. In Cheminformatics in Drug Discovery, Oprea, T.I. (Ed.). Wiley-VCH, New York, **2005**, 333-379
59 Congreve, M., Carr, R., Murray, C. and Jhoti, H. A 'Rule of Three' for fragment-based lead discovery? *Drug Discov. Today* **2003**, *8*, 876-877
60 Oprea, T.I., Blaney, J. Cheminformatics Approaches to Fragment-based Lead Discovery. In Fragment-based Approaches in Drug Discovery, Jahnke, W. and Erlanson, D.A. (Eds.). Wiley-VCH, New York, **2006**, 99-121
61 Povolna, V., Dixon, S. and Weininger, D. CABINET – Chemistry and biological informatics network. In Cheminformatics in Drug Discovery, Oprea, T.I. (Ed.). Wiley-VCH, New York, **2005**, 241-269
62 CABINET is available from Metaphorics LLC, Santa Fe, NM, http://cabinet.metaphorics.com/
63 Hansch, C., Hoekman, D., Leo, A., Weininger, D. and Selassie, C.D. C-QSAR database. Available from the BioByte Corporation, Claremont, CA, http://www.biobyte.com
64 Furet, P., Gay, B., Caravatti, G., Garcia-Echeverria, C., Rahuel, J., Schoepfer, J. and Fretz, H. Structure-based design and synthesis of high affinity tripeptide ligands of the Grb2-SH2 domain. *J. Med. Chem.* **1998**, *41*, 3442-3449